# DEEP REINFORCED KANTIAN MACHINE

BORIS D. GROZDANOFF

*Institute of Philosophy and Sociology, Bulgarian Academy of Sciences*

*Defence and International Security Institute (DISI)*

*QAISEC*

grozdanoff@gmail.com

**Abstract**

Within the diverse and rich debate around the so called "Ethical Artificial Intelligence" one of the major setbacks is the accepted lack of a unified ethical system that can be used for modelling the behavior of AI run systems like artificial moral advisors (Savulescu and Maslen 2015), autonomous vehicles, autonomous weapons and robots. Following upon an earlier suggestion for an artificial moral agent (2020) I will suggest a more general method for the quantification of key ethical concepts like consequence, behavior, rule, justification, state and property. The method would allow to harness the enormous computational power of modern computers for the purposes of in-depth ethical analysis of normative non-trivial cases. If successful it would provide a transmission mechanism, susceptible to rational evaluation, between ethics as a traditional discipline from the field of humanities and computation, as a mathematical structure of the kind found in natural sciences, where it provides not merely means of rigorous formulation but also of testing, evaluation, proofs and growth of scientific knowledge. The method introduces numerical weights assignment on the basis of ethical non-identity. Thus, if between a certain ethical dilemma with two possible actions we observe that the actions are normatively non-identical we can continue and capture this non-identity in graph-numerical terms thus effectively quantifying the normative setting. This quantification would provide us with the rigorously formalized content that would allow us to transfer this rigor onto a subsequent logical reasoning, enabling inferences, preference, justification and choice.

**Keywords:** ethical artificial intelligence, AI, ethics, artificial moral agent, ethical consequences, graphs, justification

In one of the very few practical proposals for an architecture of an artificial ethical system, Powers suggests that Kant's influential categorical imperative (CI) dictum "Act only according to that maxim whereby you can at the same time will that it should become a universal law"[1] could serve as a prospective ethical source. The main reason for this choice is Powers's interpretation of the CI as "a procedure … for generating the rules of action ... that works in a purely formal manner."[2] But there is a second reason, more general in nature, that he states

further: "Rule-based ethical theories like Immanuel Kant's appear to be promising for machine ethics because they offer a computational structure for judgment"[3], and also, "A rule-based ethical theory is a good candidate for the practical reasoning of machine ethics because it generates duties or rules for action, and rules are (for the most part) computationally tractable.[4]

Both of these are, surprisingly, not entirely in tune with Powers' own explicit tenet to harness the *logical potential* of the imperative as well as its *computational potential,* the latter being heavily understressed in his paper. Powers develops a detailed proposal of what he dubs a "Kantian machine", using his own original "reformulation of Kant's ethics for the purposes of machine ethics".[5] The proposal promises to suggest "what computational structures such a view would require …", but is nevertheless developed *entirely as a logical model,* structured around three views of how the CI works. Powers focuses on three logical aspects of (implementation of) the CI for the adapted purposes of machine ethics: consistency, commonsense practical reasoning, and coherency.[6] The logical modeling of a CI implementation is certainly natural within the field of analytic ethics and Powers makes a convincing case of the model employing deductions of actions from facts in order *to explore consistency,* recent work in non-monotonic logic and "commonsense reasoning" in order *to explore commonsense practical reasoning,* and the assumption that "ethical deliberation follows a logic similar to belief revision" in order *to explore coherency*.[7]

The model, bearing the informative name "machine", could obviously have nothing but computers or computer-operated machines in mind. Yet, those remain just cursorily marked in the fully undeveloped *promise for a computational structure*. And as Powers himself quite correctly stresses, at least a part of the very reason for choosing Kant's rule-based ethical theory, participating here in the *modi* of the CI, is the *computational tractability of rules*, on the one hand, and, on the other, its alleged ability to provide a "computational structure of judgment".[8] In the model, however, we see solely *logical tractability of the rules* and (some) *logical structure of judgment*. What we do not see is any computational structure: neither in the tractability of the rules, nor in the structure of judgments, let alone in a programming implementation of the model into a code that would eventually harness the computing power of computers for the functional purposes of the model, which are essentially logical. The latter

lack is certainly unfairly mentioned here, as a programming implementation would certainly require a separate project falling outside the scope of Power's paper.

As both philosophers (especially philosophers of mathematics and of language) and computer scientists alike are painfully aware, the computability potential that lies tacit and completely unpacked in the paper needs to be made explicit even if we want to see only a proof of concept of the model functioning outside the paper and on bare metal computing machines. The distinction between assuming (or hoping) that there might be a computable model based on this one or another similar notion of a Kantian machine, and actually demonstrating that there is one, is anything but trivial. The computational expression of any logical model, in ethics or anywhere else in logic, is also highly non-trivial, even if only partially achievable and applicable to a very limited scope of tasks.[9]

Here I would like to suggest, broadly along the lines of Power's interpretation of Kant's CI as computationally promising, another general model of a *semi-Kantian machine*. This model is, once again, based on a reformulation of the CI, but instead of focusing on logic as an instrument for developing rules of ethical behavior and examining ethical consequences, *suggests a purely computational structure as the foundation to achieve these goals*. I do agree fully with Powers that rule-based ethical theories are computationally very promising, but I believe that we can model segments of those theories, such as CI, directly via computational models and not by introducing the logical model-in-the-middle. I am practically skeptical regarding direct logical implementations of artificial ethical models, like Powers' Kantian machines, on the one hand due to the uncertainty of their expected ethical success, and on the other hand – due to the highly non-trivial technical task of encoding such models as computational models, which both works and preserves the desired logical functionality of a true Kantian machine in the spirits of Powers's interpretation of Kant's CI. Suggesting direct logical models is reminiscent of the old approach of expert systems, while today the AI architects have at their disposal the immense power of technologies like artificial neural networks (ANN) and reinforcement learning (RL). Those, however, are purely computational, of course, and any attempt to build a logic-based Kantian machine would unnecessarily burden them with a model-in-the-middle, since any machine implementation of a logical model would need a computational engine in order to function at all; and ANN and RL are two of the leading choices of architecture for such engines.

**Deep Reinforcement Learning**

One of the most successful AI technologies today is the subspecies of the Reinforcement Learning (RL) called Deep Reinforcement Learning (DRL). RL analyzes a task in terms of its components, e. g. agent, environment, observation, state, reward and policy via what computer science and pure mathematics refer to as Markov Decision Process (MDP).[10]

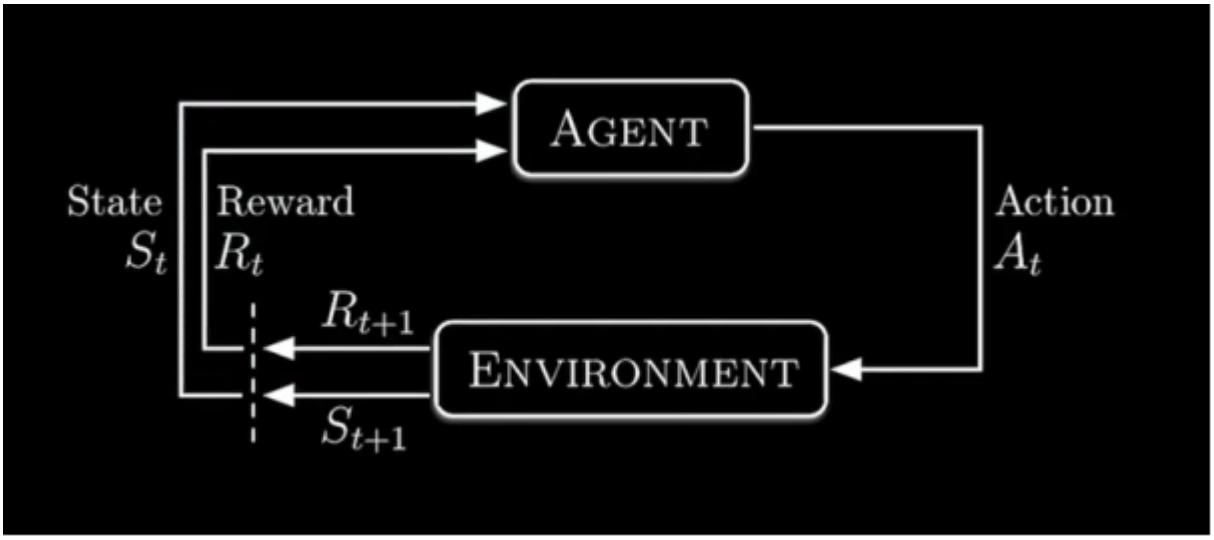


*Fig. 1 The Reinforcement Learning Cycle [11]*

The agent, as the name suggests, is the acting module, which is the subject of the RL artificial learning task; he observes the states of the environment he is in, performs a purely computational "analysis" and *chooses* to perform an action that leads to a change in the environment, which, in turn, is the next state the agent would observe in it. The agent is governed, on the one hand, by the observed state of the environment, and on the other hand – by the "analysis" and the reward signal that he receives as a result of his action. Based on the magnitude of the rewards, the agent gradually, along the course of many iterations of this process, *learns* to act in such a way as to maximize the expected future sum of the rewards. The environment is the general setting which provides the "world" of the agent and contains the allowed actions; it renders the reward and eventually includes other agents that act from within the environment, typically against the leading agent who learns. Those adversarial agents are observed by the leading agent as part of the environment and their actions against the "goals"

of the agent are key for a multitude of learning processes, especially in adversarial games like Chess, Go, Dota 2, the Atari type games and others.

The key mathematical computation regulating the agent's "decision" to take a certain next action is the probability function that computes *the probability of the next state* of the environment and *the probability of the next expected reward*, given some concrete values of the *current state* of the environment observed by the agent and the agent's *current action*. This function is called a distribution function and usually looks like this:

$$p(s', r|s, a) = \text{Prob}(S_{t+1} = s', R_{t+1} = r \,|\, S_t = s, A_t = a)$$

*Fig. 2 The Distribution function*

Without going into further technical details about RL modeling, it is sufficient for our present purposes to stress that the "Deep" sub-species of the model family harnesses the learning power of artificial neural networks (ANNs) by integrating two (or one common, depending on the type of the model) neural networks, which discover both *the optimal policy* of agent behavior required in order to win the game and *the optimal value function*.

**Gamifying Ethical Behaviour**

It is certainly not far from the truth that Deep Reinforcement Learning can be considered among the most successful architectures for AI, if not outright the most successful one. Among its most notable successes are DRL model wins in the games of Chess, Go, DOTA 2 and a large number of ATARI games. DRL also gained momentum in other fields which are quite distinct from entertaining and intellectually stimulating games, but still share a key feature with them: warfare, pilot training, wargaming, swarm systems (like drones), on-demand logistics, fintech (optimized trade execution, portfolio optimization, market making) and others. What is common for all of those, as well as many other real world high-stake tasks, *is the possibility to model the tasks in a game-like setting*. Thus, they are naturally susceptible to deep reinforcement learning modeling. So far no DRL model has been proposed or developed that has produced what Savulescu and Maslen call an artificial moral agent (AMA). Due to the enormous potential of DRL and due to the nature of human ethical behavior, which can be

modeled by standard DRL modules (like agent, environment, observation, action and reward), we are justified in expecting that using a DRL approach in order to create an AMA has great potential.

The key task in the DRL AMA model would be to represent the participating ethical elements as an integrated part of the agent-environment interaction. The ethical task, as far our AMA is concerned, would be to teach the agent to behave ethically in a normatively relevant environment where he learns to do so by observing the states of the environment, choosing his own actions that change the environment, and receiving rewards. The choices the agent makes are choices of action where he has no control over the states of the environment (apart from his action introducing a next change in the environment). In these choices, the DRL agent is driven by the reward as he figures out which action in a particular state of the environment provides the maximum sum of future rewards by *exploring* the actions and *exploiting* the actions; these are RL' main methods of navigating the space of behavior choices.

From an ethical point of view, the key challenge in this model is *what to introduce as ethical behavior* and how this could be DRL modeled in order to be learned by the agent. After all, ethical debates about the nature of good and evil are some of the oldest and most complex ones, and remain unresolved to this day. I propose that for the concrete purposes we use a bottom-up approach regarding the ethical behavior modeling, where we introduce a hard-to-dispute-as-unethical behavior, on the one hand, and hard-to-dispute-as-ethical behaviors, on the other, as elements in the model. Also, we can do so for a limited number of behaviors, thus *diminishing the complexity of the task to a practical computability*, since an increasingly complex ethical model attempting to approximate real-life ethical scenarios would require enormous computing power available only to the leading giants in AI, such as Deep Mind and Open AI.

We can begin by introducing the following small number of ethically clear behaviors:

| | *Actions* |
|---|---|
| *Ethical* | 1. Tell truth<br>2. Protect from hit<br>3. Protect from torture |

| | |
|---|---|
| | 4. Save from kill<br>5. Permit good options (1-4) by inaction |
| *Unethical* | 6. Lie<br>7. Hit<br>8. Torture<br>9. Kill<br>10. Permit bad action (6-9) by inaction |

Those behaviors would be introduced in the model as forming the space of actions ($S_a$) available to the agent. The task before the agent, in order to "win the ethical game", is to learn along the course of his interactions with the environment and the other agents how to behave ethically. The four kinds of behavior-actions, *actively good, actively bad, passively good* and *passively bad*, represent the typical set of real-life ethical scenarios as closely as practically possible, even if not exhaustively. Thus, the DRL model would be a multi-agent model where the ethical actions available to the other agents would be:

| | Good Agent | Bad Agent | Mixed Agent | Inactive agent |
|---|---|---|---|---|
| Action sets | All his actions are good (1-5) | All his actions are bad (6-10) | All actions are available to him (1-10) | Acts only passively, both good passive and bad passive (5 & 10) |

The ethical nature of the action is predefined in the model for each agent in the environment, as the learning agent is ignorant of both their nature and any differences between them. Reward values would initially be:

| Actions | Reward points |
|---|---|
| 1. Tell truth | + 5 |
| 2. Protect from hit | + 10 |

| | |
|---|---|
| 3. Protect from torture | + 50 |
| 4. Save from kill | + 100 |
| 5. Permit good options (1-4) by inaction | + 1 |
| 6. Lie | - 5 |
| 7. Hit | - 10 |
| 8. Torture | - 50 |
| 9. Kill | - 100 |
| 10. Permit bad action (6-9) by inaction | - 5 |

These reward points are only initial and are only applied during the learning process. During training, the AI architect would gradually fine-tune them in such a way as to reach the set of reward points that delivers the optimal policies of action behavior of the agent, as well as the optimal value function. Multiple reward points assignments would be run and game-success parameters would be assessed by the DRL architect in order to see which reward point distribution leads to whose win (that of the ethical agent, of others, or of the "game"); how fast the win is attained, and *which wins have minimal ethically negative events* (hits, kills) while also having the fastest win and the highest number of ethically positive events (protections, etc.).

The practical proposal here, limited as it is in terms of ethical power, nevertheless has the potential to model a number of difficult-to-dispute-as-(un)ethical behaviors. The environment is *fully normative, that is, every action has a deterministic normative weight*, and every effect of an action, modeled by the next state of the environment, can be traced back to a single action or a preceding chain of actions.

By following Powers's interpretation of Kant's categorical imperative as a procedure for generating rules for action, the model proposes an initial and as close as possible to reality ethical assessment that relates to both *consequentialist* and *deontological interpretations*. Here the consequentialist interpretation, which generally judges an action by its effect's assigned ethical value, is quantified by a weight arrived at by the DRL model, and represented through the *possibility to assess the ethical consequences of the agent's actions and, ultimately, the*

*agent's total behavior,* available as a separate ethical game-play or as a total game episode. The model provides a means to distinguish between separate, *discrete ethical actions* and *a "chained" strategy of ethical behavior*, oriented toward a concrete ethically assessable goal such as the win of the episode or the total game, which consists of a predetermined number of episodes, ideally equal to the minimal number of games necessary for the agent to learn the most ethical behavior within this limited setting.

The deontological interpretations of ethics are represented through the *model's ability to generate rules for action* along the lines of Powers's general suggestion of Kantian machines, but, again, entirely computational and not logical. It also, upon successful training, would provide *a computational justification for assessing agents as good and bad, as well as their ethical traits* modeled by policies and strategies of behavior following the deontological ethical tradition. What might be of theoretical interest here for ethicists is that the model blends elements of both consequentialist and deontological approaches to ethics in such a way that they depend on each other and do not collide in the way in which they usually appear to do in the literature. If successful, the DRL model would have the theoretical benefit of showing that *both traditions are actually complementary instead of contradictory*.

The world of the agent or the world of this ethical "game" is certainly limited, but nevertheless, combinatorially very rich: it contains, as rules of the game, 10 allowed types of actions, and it figures multiplayer complexity with 4 separate agents, each with its own ethical-behavioral characteristics. This setting approximates *being exhaustive for modeling types of ethical behaviors* found in the real world, where we typically have those 4 types of ethical agents. While, of course, the reality would be that perhaps over 99% of ethical agents in the real world are of the mixed type, as far as there are virtually no purely evil agents and no purely good or purely inactive agents. The actions of the model are defined as contrast on purpose, in order to diminish their susceptibility to be disputed from a number of ethical-theoretical stances. Thus, not many ethicists would a priori assess actions 1-4 as anything other than being good actions in themselves, even if they can lead to bad effects (if the agent saves a baby-Hitler, to give a classic example). Obviously, *the normative classification of the allowed actions pertains to the actions only and it does not necessarily transfer onto their consequences*. Good actions can lead to bad consequences and bad actions can lead to good consequences (killing a serial killer, to give another classic example), and yet the action, taken by itself, before producing any

consequence, does have its own normative weight and this weight is reflected by the table of initial rewards above.

The model allows for all possible combinations and thus is *quite rich in terms of ethical modeling power*. It provides quantifiable, traceable, justifiable means to assess actions and consequences as good or bad, as well as means to assess behavior strategies as good or bad; based on all this, it can provide *a quantified normative justification to assess a person* (here modeled by an agent) *as good or bad*. It can also predict, upon training, future consequences of certain separate or bundled behaviors, including those emerging in training normative behavior strategies (of the kind "in X type scenarios, do Y").

An expected benefit of the model is its ability to be quickly modified, in terms of both structure and complexity. Thus, we can reproduce re-scaled versions of it with a much larger and very different setting: with only bad agents, with only good agents, with only mixed agents (as this latter version would be much closer to real life scenarios). It can grow in sheer complexity, for example in a setting with 100 or 1000 multi-agents embedded in the environment (all agents in the DRL model are regarded by the learning agent as parts of the environment). Another expected potential benefit is *the possibility to trace back* and *explain winning normative strategies,* and assess them by comparison. Thus, meta-strategies would emerge, with each ethical winner winning in its own game-setting. Due to the demonstrated potential of successful DRL models to generalize and perform in a manner superior to novel scenarios, a well-trained model would be a prominent candidate for justified normative behavior in novel scenarios. Such scenarios might include best known options, comparability to human behavior, or zero knowledge options where the agent is forced to act ethically without previous training.

The model can also include (as agents embedded in the normatively relevant environment) real-world human behavior. This can be achieved via training adversary agents in a manner that approximates actual human (un)ethical behavior, which would allow for the model to be virtually deployed to real-life ethical scenarios without actually being a real-life ethical agent among unsuspecting (or test group) humans.

In summary, an optimized and sufficiently powerful version of this model can potentially become the best "ethical game" in the digital domain, where both AI architects and professional ethicists might create a "world champion", not unlike the famous occasions when

DRL models such as AlphaGo,[12] AlphaZero[13] and MuZero[14] became world champions in the games of Chess and Go.

An important and likely objection needs to be addressed here. It usually goes something like this: since the discipline of ethics has no universally agreed upon ethical system[15], there is no way we can teach an AI agent how to be an AMA, or artificial moral advisor, let alone how to be an artificial moral agent. The DRL model suggested here can help us answer this objection in the following way. First, the fact that there is no unified and universally agreed upon ethical system does not affect the real-life behavior of humans. Thus, even if we eventually arrive at such a theory, it would perhaps take some time for it produce a significant impact on human ethical behavior – that is, if it ever produces this impact at all. This fact notwithstanding, we face a very different case with AI-run machines: states, industry and societies implement them at an increasingly fast pace and they become *a de facto ethical factor in our lives.* The dilemma which humanity must resolve as quickly as possible, as it faces (un)ethical behavior from already operational AI technologies, is the following: to either continue delving deep within ethics as a theoretical discipline with a rich, millennia-long controversial history, or attempt to harness the computational power of the very same AI technology in order to see whether it can arrive at superior ethical solutions while humanity still ponders a number of contradictory theories.

The DRL model, once operational, will at best shed light on actual ethical problems and perhaps enable optimal ethical behavior, thereby winning the competition against human ethics; or, at worst, it will provide a novel and yet more rigorous ethical theory. The only way to see how these scenarios might play out is to build, train, analyze and deploy on novel (virtual) data this and other, alternative, AI ethical models.

**NOTES**

[1] See *Grounding of the Metaphysics of Morals*.

[2] Powers, op. cit., p. 46, omission is mine.

[3] Ibid.

[4] Ibid, p. 46–47.

[5] Ibid, p. 47.

[6] Ibid.

[7] Ibid, p. 47.

[8] Ibid, p. 46.

[9] See the demise of the logical expert systems, for more optimizing results see the CYC Project and the modern hybrid neural-symbolic models.

[10] For an in-depth explanation of Markov Decision Process see https://towardsdatascience.com/introduction-to-reinforcement-learning-markov-decision-process-44c533ebf8da.

[11] Figures 1 and 2 are taken from the formalization of the DRL online course of Duane Rich. The course can be found online at: https://www.youtube.com/@Mutual_Information.

[12] Silver, D., Huang, A., Maddison, C. et al. (2016). Mastering the game of Go with deep neural networks and tree search. // *Nature*. Vol. 529, pp. 484–489; https://doi.org/10.1038/nature16961.

[13] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, Demis Hassabis, (2017) «Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm», arXiv:1712.01815 [cs.AI]

[14] Schrittwieser, J., Antonoglou, I., Hubert, T. et al. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. // *Nature*. Vol. 588, pp. 604–609; https://doi.org/10.1038/s41586-020-03051-4.

[15] See Serafimova (2020).

## BIBLIOGRAPHY

**Allen, C., Varner, G. and Zinser, J.** (2000). Prolegomena to any future artificial moral agent. // *Journal of Experimental & Theoretical Artificial Intelligence*. Vol. 12, Issue 3, pp. 251–261.

**Anderson, S.** (2011). How machines might help us to achieve breakthroughs in ethical theory and inspire us to behave better. // Anderson, M., Anderson, S. (Eds.), *Machine Ethics*. New York: Cambridge University Press, pp. 524–530.

**Anderson, M., Anderson, S.** (2007). Machine ethics: Creating an ethical intelligent agent. // *AI Magazine.* Vol. 28, Issue 4, pp. 15–26.

**Beavers, A.** (2001). Kant and the problem of ethical metaphysics. // *Philosophy in the Contemporary World*. Vol. 7, Issue 2, pp. 47–56.

**Beavers, A.** (Ed.). (2010). Robot ethics and human ethics. // *Special issue of Ethics and Information Technology*. Vol. 12, Issue 3.

**Floridi, L., Sanders, J.** (2004). On the morality of artificial agents. // *Minds and Machines*. Vol. 14, Issue 3, pp. 349–379.

**Gips, J.** (1995). Towards the ethical robot. // Ford, K., Glymour, C., Hayes, P. (Eds.). *Android Epistemtology.* Cambridge, MA: MIT Press, pp. 243–252.

**Grozdanoff, B. D.** (2020). Prospects for a Computational Approach to Savulescu's Artificial Moral Advisor. // *Ethical Studies.* Vol. 5, Book 3, 2020, pp. 121–140.

**Grozdanoff, B. D.** (2021). Graph Formalism for normative ethical tasks in Artificial Intelligence. // *Ethical Studies.* Vol. 6, Book 2, 2021, pp. 109–120.

**Kant, I.** [1785] (1981). *Grounding for the Metaphysics of Morals*. Indianapolis, IN: Hackett Publishing Company.

**Kurzweil , R**. (1990). *The Age of Intelligent Machines*. Cambridge, MA: MIT Press.

**Mittelstadt et al.** (2018). The ethics of algorithms: Mapping the debate. // *Big Data and Society*. Vol. 3 Issue 2; https://doi.org/10.1177/2053951716679679

**Moor , J.** (2006). The nature, importance, and difficulty of machine ethics. // *IEEE Intelligent Systems 1541–1672*, pp. 18–21.

**Powers , T.** (2006). Prospects for a Kantian machine. // *IEEE Intelligent Systems 1541 – 1672*, pp. 46 – 51.

**Savulescu J., Maslen H.** (2015). Moral Enhancement and Artificial Intelligence: Moral AI?. // **Romportl J., Zackova E., Kelemen J.** (Eds.). *Beyond Artificial Intelligence. Topics in Intelligent Engineering and Informatics.* Vol 9, Springer, Cham; https://doi.org/10.1007/978-3-319-09668-1_6

**Schrittwieser, J. et al.** (2020). Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. // Deepmind.

**Serafimova, S.** (2020). *Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement.* // *Humanit Soc Sci Commun.* Vol. 7, p. 119: https://doi.org/10.1057/s41599-020-00614-8

**Silver, D. et al.** (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. // Deepmind.

**Sutton, R., Barto, A.** (2018). *Reinforcement learning: An Introduction (2nd Ed)*. MIT Press.

**Todorova, M.** (2020). *The Artificial Intelligence: A Brief history of its development and the ethical aspects of the problem*. Sofia: East-West Publishing, in Bulgarian.

**Turing, A.** (1950). Computing machinery and intelligence. // *Mind*. Vol. 59, Issue 236, pp. 433–460.